Journal international des technologies de l'information hybrides
Vol.8, n°8 (2015), pages 279-288
http://dx.doi.org/10.14257/ijhit.2015.8.8.29

Reconnaissance d'entités nommées biomédicales basée sur une neutralité profonde Réseau

Lin Yao1,2,3 , Hong Liu1, Yi Liu2 , Xinxin Li4 et Muhammad Waqas Anwar5

1École d'ingénierie électronique et d'informatique, Université de Pékin
2Établissement Pku-hkust de Shenzhen-hong Kong
3École de logiciel, HIT
4Département d'informatique, HITSGS
5Département d'informatique, Institut de technologie de l'information COMSATS
yaolin@insun.hit.edu.cn

Abstrait

De nombreuses méthodes d'apprentissage automatique ont été appliquées à la reconnaissance d'entités nommées biomédicales et obtiennent de bons résultats sur le corpus GENIA. Cependant, la plupart de ces méthodes répondent à l'ingénierie des fonctionnalités qui demande beaucoup de main d'œuvre. Dans cet article, d'énormes informations sur les caractéristiques potentielles représentées sous forme de vecteurs de mots sont générées par des réseaux neutres basés sur des fichiers texte biomédicaux non étiquetés. Nous proposons une méthode de reconnaissance d'entités nommées biomédicales (Bio-NER) basée sur une architecture de réseau neuronal profond qui comporte plusieurs couches et chaque couche résume les caractéristiques en fonction des caractéristiques générées par les couches inférieures. Notre système a obtenu un score F de 71,01 % sur le corpus de test régulier GENIA, les valeurs du score F pour une validation croisée 5 fois sont de 71,01 % et ce résultat est proche des performances de pointe avec uniquement POS (partie de -discours) et représente l'apprentissage en profondeur qui peut être effectué efficacement sur le NER biomédical.

Mots clés : Deep learning, Reconnaissance d'entités nommées biomédicales, Réseaux neutres

1. Introduction

Dans le domaine biomédical, des données au niveau du gigaoctet, voire du téraoctet, sont produites chaque jour. Des données aussi énormes stimulent le développement de la recherche dans le domaine biomédical à bien des égards. Cependant, la capacité des chercheurs biomédicaux à analyser et à appliquer le Big Data biomédical est souvent limitée. Des logiciels et outils pertinents pour le traitement de l'information biomédicale pourraient améliorer l'utilisation de ces énormes données. Le NER biomédical est une étape initiale cruciale pour le traitement de l'information biomédicale. Et c'est une technologie fondamentale pour identifier les entités et leurs interactions. Mais le NER biomédical est plus difficile que le NER général en raison de situations complexes telles qu'une expression irrégulière, des frontières difficilement distinguées et des membres du groupe qui changent quotidiennement. La difficulté et l'importance potentielle de cette tâche attirent de nombreux chercheurs [1-5].

De nombreuses techniques d'apprentissage supervisé ont été utilisées pour résoudre le problème biomédical du NER, telles que les HMM (Hidden Markov Models) [6], les MEMM (Maximum Entropy Markov Models) [7, 8], SVM (Support Vector Machines) [9, 10], et CRF (Conditional Random Field) [11-13]. Le CRF est appliqué pour aborder la reconnaissance d'entités dans le domaine biomédical par Settles [9]. La méthode atteint un F-score de 69,9% sur le corpus GENIA avec seulement plusieurs types de fonctionnalités. Le HMM est porté sur le corpus GENIA et atteint une précision de 66,5% et un rappel de 66,6% [9]. SVM est appliqué par Ki-Joong Lee pour présenter un système de reconnaissance d'entités nommées en deux phases sur le corpus GENIA et obtenir un score F de 74,8 % pour l'identification des limites et un score F de 66,7 % pour la classification sémantique [10].
Le modèle de champs aléatoires conditionnels (CRF) à saut de chaîne a été appliqué à cette tâche en considérant les dépendances à longue portée. Cette approche atteint un F-score de 73,2% sur GENIA

ISSN : 1738-9968 IJHIT

corpus[14]. Li présente un modèle Bio-NER en deux phases sur le corpus GENIA. Ils ont divisé la tâche en deux étapes : la détection d'entités nommées (NED) et la classification d'entités nommées (NEC)[15]. La première étape consiste à distinguer les entités non nommées (NNE), mais sans identifier leurs types. Six classificateurs sont construits dans cette phrase. Dans l'expression NEC, la stratégie multi-agents est utilisée et ils obtiennent un score F de 76,06 %. Cependant, les méthodes précédentes sont soit basées sur des modèles de fonctionnalités spécifiques créés à la main, soit sur une méthode d'empilement intégrée à différentes méthodes de formation. La création de modèles de fonctionnalités est un processus entièrement empirique et nécessite de nombreux travaux d'essai.

Motivés par les travaux de Collobert [16] , nous construisons un modèle de réseau neuronal pour la tâche biomédicale de reconnaissance d'entités nommées. Nos travaux montrent que l'apprentissage profond peut être effectué efficacement sur le NER biomédical. Notre architecture a atteint des performances proches de l'état de l'art sur le corpus GENIA, qui est un corpus standard populaire, et a été adoptée par de nombreux groupes de recherche à des fins d'évaluation.

Cet article est organisé comme suit. La section 2 décrit l'architecture proposée basée sur un réseau neutre profond. Dans la section 3, nous présentons notre performance suivie de notre analyse. Les conclusions sont formulées dans la section 4.

2. Architecture

Notre méthode est basée sur un CNN (convolutional neural network), qui a été adopté avec succès dans certaines tâches NLP (Natural Language Processing) [17-19]. Les architectures polyvalentes de réseaux neuronaux convolutifs ont été proposées par Bengio [17] pour le modèle de langage probabiliste. Depuis lors, les NN ont été réintroduits plus tard pour plusieurs tâches de PNL. Nous le choisissons pour la tâche Bio-NER. L'architecture du réseau neuronal est illustrée à la figure 1.

De nombreuses approches de marquage traditionnelles doivent choisir différentes fonctionnalités en fonction de différentes tâches et de nombreuses connaissances préalables et professionnelles. L'ingénierie des fonctionnalités est importante mais manuelle et demande beaucoup de travail. Par rapport à d'autres systèmes de sur-ingénierie, l'approche d'apprentissage en profondeur réduit la dépendance à l'ingéniosité linguistique. La figure 1 décrit le mot bêta qui se trouve au milieu droit de la fenêtre glissante et qui est évalué à l'instant t. L'entrée de notre modèle de réseau neuronal correspond aux mots dans les fenêtres glissantes représentés sous forme de vecteurs à valeur réelle. Après transformation des couches linéaires et de la couche sigmoïde, le score de nœud pour chaque étiquette du mot bêta est généré. Comme le montre la dernière procédure de la figure 1, un réseau de scores pour une phrase est finalement généré. Les nœuds de chaque colonne indiquent les scores des étiquettes à un moment donné. Les scores de transition des étiquettes sont répertoriés sur le bord. Viterbe L'algorithme est appliqué pour obtenir la séquence d'étiquettes optimale.

2.1. Extraction de vecteurs de caractéristiques de mots

La longueur d'entrée de CNN est fixe et doit être adaptée aux données texte. Comme le montre la figure 1, tout d'abord, un dictionnaire de mots W sera construit à partir d'énormes données brutes provenant d'articles biomédicaux. Chaque mot du dictionnaire est représenté par un vecteur de dimension fixe. Les mots sont transformés en vecteurs pour l'entrée CNN. Les vecteurs représentent les caractéristiques transportant des informations sur les phrases et les similitudes sémantiques entre les mots. Les représentations vectorielles des mots sont stockées dans la matrice $M \in \mathbb{R}^{D \times |W|}$, où D est la dimensionnalité vectorielle pour représenter un mot, et | | W | est la taille du vocabulaire. Habituellement, nous considérons le vocabulaire des mots comme fini. Le fichier vectoriel de mots résultant peut également être utilisé comme fonctionnalités dans d'autres applications de traitement d'informations biomédicales et d'apprentissage automatique. Les paramètres M sont initialisés de manière aléatoire et formés sur une grande quantité de fichiers papier texte biomédicaux non étiquetés avec un réseau neuronal général.

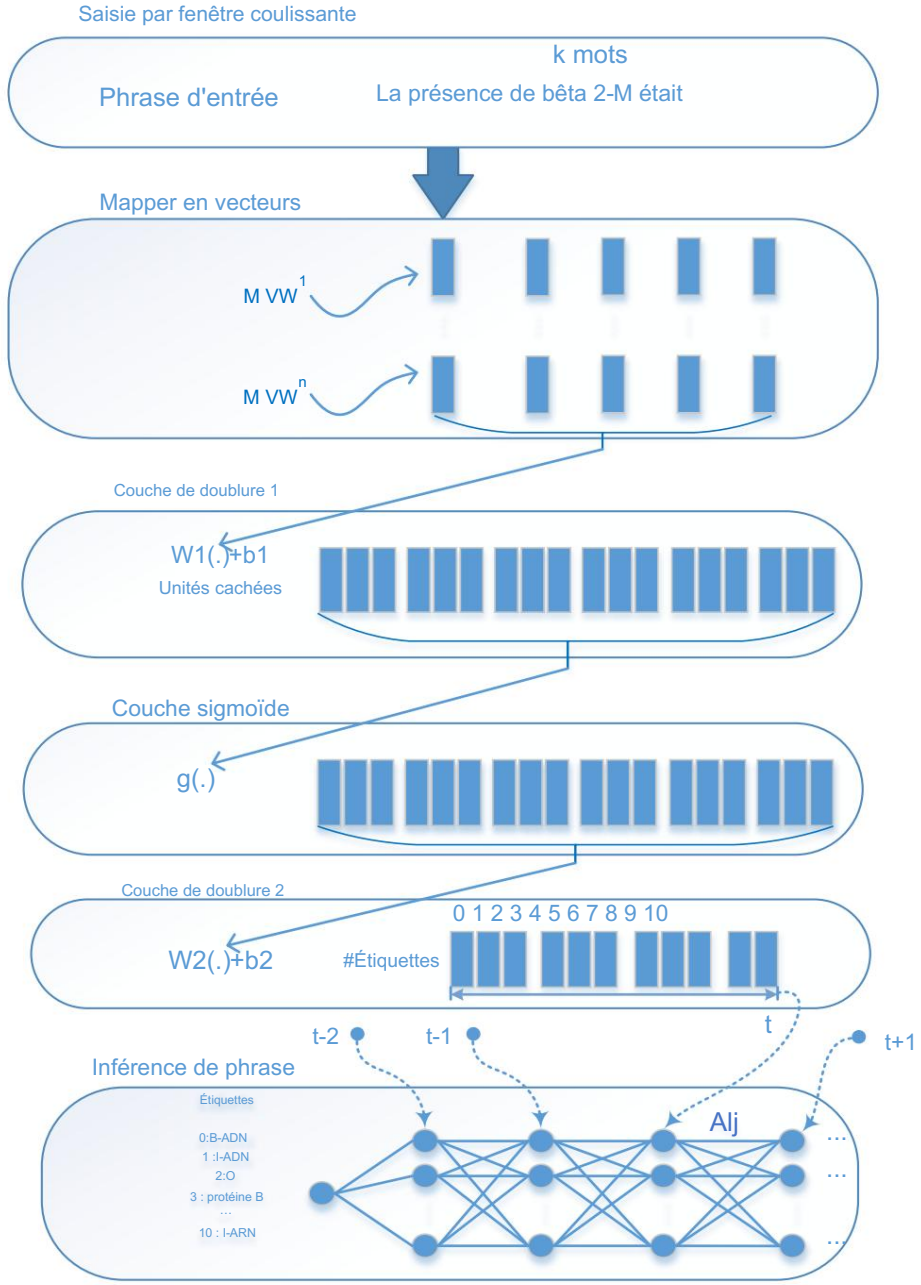

Figure 1. L'architecture du réseau neuronal

Généralement, la représentation vectorielle à valeur réelle M peut être obtenue de deux manières. Un initialise le vecteur pour chaque mot i avec zéro dans toutes les positions sauf 1 pour la position Mi , et les optimise automatiquement en tant que paramètres pendant la phase d'entraînement du réseau [18]. La deuxième façon consiste à considérer la représentation des mots comme faisant partie de la formation d'un modèle de langage de réseau neuronal [20-23]. Dans cet article, nous les optimisons sur une tâche spécifique Bio-NER. Dans cet article, nous adoptons la seconde. Comparé à différents modèles de langage [23-26], nous choisissons le modèle de langage du réseau neuronal skip-gram qui n'est pas le plus rapide mais plus adapté à l'entraînement des mots rares, car de nombreux mots rares n'apparaissent que dans les littératures biomédicales.

De plus, si des fonctionnalités autres que des mots peuvent être apportées à la tâche, elles peuvent être ajoutées en tant que dimension des vecteurs de mots. Par exemple, une fonctionnalité peut être utilisée pour indiquer si un mot est un mot temporel ou s'il est majuscule. Dans nos travaux, pour vérifier l'affection de NN profond

sur Bio-NER, une seule fonctionnalité est concernée. Nous le décrirons plus tard.

2.2. Extraction des caractéristiques du niveau de phrase

Pour la tâche NER biomédicale, chaque mot dans les phrases doit recevoir une étiquette appropriée pour indiquer s'il s'agit d'une NER biomédicale ou non. Les phrases seront les entrées du système. Les sorties sont des séquences d'étiquettes appropriées pour chaque phrase. Étant donné que la longueur des phrases est variable, mais que la saisie de NN est fixe, nous choisissons l'approche par fenêtre glissante comme beaucoup d'autres problèmes d'apprentissage automatique. La taille de la fenêtre est k , qui est décidée au début et le système peut avoir une précision différente à cause de cela. En utilisant l'approche par fenêtre glissante, les informations de dépendance entre l'étiquette de chaque mot et ses mots voisins sont concernées. Ses voisins dans la fenêtre traverseront donc ensemble le calque. Lorsque le mot en position p est étudié, lui et ses voisins de positions comprises dans l'intervalle de

(. pk 1) / 2,(pqt . 1) / 2 seront transmis à la couche Mapping. Comme nous l'avons mentionné précédemment, chaque mot a été traduit en un vecteur dimensionnel via cette couche. Par conséquent, la taille d'entrée de la couche linéaire 1 est fixée à D k .

2.3. Critère d'étiquette

Un réseau de neurones profond peut être représenté comme une architecture à plusieurs couches. Chaque les fonctionnalités abstraites des couches sont basées sur les fonctionnalités générées par les couches inférieures. Dépendant du conception, chaque couche peut être soit une fonction linéaire, soit une autre transformation. Une fonction f , décrivant les trois couches de notre architecture peut être notée comme suit :

$$f_x(M) = g(xb^2 + (M^1 \cdot \cdot \cdot^1)^2 \quad (2.1)$$

où les matrices $M^{1\ H\ NSP}$, b^1 $^{1\ heure}$, M^2 $^{|\ LH}$, b^2 $^{1\ |\ L}$, g(.) est sigmoïde fonction. H est le nombre d'unités cachées et peut être ajusté pour de meilleures performances. L est la taille de tous les jeux d'étiquettes d'étiquettes possibles.

En utilisant l'ascension de gradient stochastique, sur un ensemble d'entraînement T, le paramètre v-dimensionnel matrice (, , ,...,) est entraîné en maximisant la vraisemblance :

$$\sum_{(x,y)\in T} \text{journal}(p(y \mid x, \theta)) \quad (2.2)$$

Cette tâche est un problème de classification multi-classes. fxl , , . est utilisé pour désigner le score de chaque lième étiquette dans l'exemple donné x qui correspond à une fenêtre de mots d'entraînement.

fxl , , . peut être interprété comme une probabilité conditionnelle p. l | x,. doit . . En appliquant le opération de régression softmax, p. l | X, être décrit comme :

$$\text{sv}(p \mid ,) \quad \frac{e^{fxl(,,)}}{\sum_j e^{fxj(,,)}} \quad (2.3)$$

Définir l'opération d'ajout de journal comme

$$\text{journal ajouter z} \quad \text{journal} \quad e^{z} \quad (2.4)$$

Par conséquent, la log-vraisemblance pour un exemple d'entraînement (,) xy peut être exprimée comme suit :

$$\text{journal}(p(y \mid x,)) = f_{x,y}(,) - \text{logadd}_j (, j,) \quad (2.5)$$

. f x[1:T], l, t, θ est défini comme le score de sortie de la phrase x[1:T] et pour le tag l à l'instant t par le système avec paramètres .

En tant que tâche d'étiquetage de séquence, le Bio-NER doit concerner le score de chaque chemin d'étiquetage. Il existe des dépendances entre ces balises dans la même phrase. Par exemple, les mots limites gauche de la catégorie d'entité A ne peuvent pas être suivis par les mots intérieurs d'une catégorie B différente. Par conséquent, pour interpréter le résultat, nous considérerons non seulement chaque mot de la phrase, mais également les dépendances entre les étiquettes. Étiqueter chaque mot de la phrase revient à rechercher un chemin avec un chemin de score maximum dans le graphique. Le score de le long du chemin dans une phrase est la somme de deux parties. L'un est les scores de nœuds mentionnés précédemment, l'autre est les scores de transition Alj qui désignent les transformations de probabilité de l'étiquette l à j. Tous les paramètres, y compris Alj et sont notés .

Pour une phrase , la partition d'un chemin avec des étiquettes est noté comme suit :

$$S(x, y_{[1:T]}, \theta) = \sum_{t=1}^{T} A_{y_{[t-1]}, y_{[t]}} \cdot f_{[1:T]}(y_t, x, t, \theta) \quad (2.6)$$

Faites les mêmes opérations le plus tôt possible. Le porter à l'exponentiel le rend positif et le respecter avec tous les chemins pour le normaliser. Par conséquent, la probabilité log-conditionnelle de prendre le vrai chemin étiqueté est interprété comme suit :

$$\log(p(y_{[1:T]} \mid x_{[1:T]}, \theta)) = S(x, y_{[1:T]}, \theta) - \log\left(\sum_{y'_{[1:T]}} \exp(S(x, y'_{[1:T]}, \theta))\right) \quad (2.7)$$

Lors de la procédure d'entraînement, tous les paramètres θ sont entraînés sur toute la durée de l'entraînement (exemples) pour maximiser le journal . Plus tard dans l'inférence procédure, l'algorithme de Viterbi est choisi pour trouver .

2.4. Dégradé stochastique

Les algorithmes de descente de gradient sont les algorithmes d'optimisation les plus simples pour minimiser une formule, mais compte tenu du coût énorme du calcul, nous choisissons une méthode d'optimisation, la descente de gradient stochastique[27]. A chaque étape d'itération, un exemple est tiré au hasard, et une nouvelle valeur de est calculé. ,

$$\theta^{t+1} = \theta^{t} - \eta_t \cdot \frac{\partial \log(p(y_{[1:T]} \mid x_{[1:T]}, \theta))}{\partial \theta^t} \quad (2.8)$$

désigne le gradient du taux d'apprentissage par rapport à et le par est une petite constante positive.

3. Expériences

3.1. Description de la tâche

La tâche du bio-ADN consiste à reconnaître les entités allant des noms de protéines/gènes aux noms de maladies/virus et à les étiqueter avec des étiquettes particulières dans un texte biomédical simple. Comme le montre la figure 2, chaque mot d'une phrase est considéré comme un jeton associé à une étiquette. Les étiquettes telles que BC, IC ou O indiquent non seulement la catégorie de l'entité nommée (NE) mais également l'emplacement du jeton au sein de l'entité nommée. Dans la dénotation de l'étiquette, C représente les étiquettes de catégorie ; B et I sont des étiquettes de localisation, représentant respectivement le début d'une entité et l'intérieur d'une entité. Il existe 5 catégories d'étiquettes : protéine, ADN, ARN, type_cellule, lignée_cellulaire. O indique qu'un jeton ne fait pas partie d'un NE. Dans le corpus GENIA, le fichier de test est donné en notation BIO. Au total, 11 étiquettes sont incluses en utilisant la notation BIO, comme le montre la figure 2. Chaque jeton du texte biomédical se verra attribuer l'une des 11 étiquettes dans les résultats de reconnaissance.

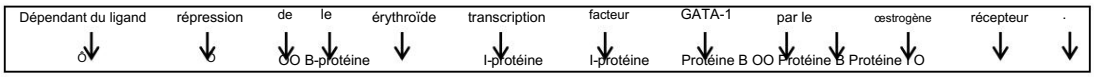

Figure 2. Exemple de reconnaissance d'entités nommées biomédicales

3.2. Résultat de l'expérience et analyse

La représentation des mots est entraînée à l'aide d'un modèle de langage de réseau neuronal skip-gram avec des données d'entraînement non étiquetées. Les données non étiquetées sont collectées à partir de la base de données PUBMED à l'aide des outils biopython1 et « médicament », « interaction », « protéine », « ADN », « type de cellule » sont choisis comme mots-clés pour la recherche. Nous téléchargeons au total 339 084 articles associés à partir de la base de données publiée et 294 993 articles contiennent des résumés. Un fichier texte de 431 Mo au total est utilisé comme ensemble de données non étiquetées. L'outil Word2vec2 est utilisé pour implémenter notre modèle de langage skipgram. Au total, 205924 mots avec 600 vecteurs de dimension sont inclus dans le dictionnaire W.

Outre ces fonctionnalités, nous considérons également la fonctionnalité POS utilisant des outils de marquage de parties du discours3 qui est spécifiquement conçue pour le texte biomédical puisque les caractéristiques du texte biomédical sont assez différentes de celles des articles de journaux.

Dans notre expérience, le corpus GENIA est appliqué. La précision, le rappel et le score F sont sélectionnés comme mesure d'évaluation. La précision est le nombre de NE détectés correctement par un système divisé par le nombre total de NE identifiés par le système. Le rappel est le nombre de NE qu'un système a correctement détecté divisé par le nombre total de NE contenus dans le texte d'entrée. F score précision* Recall / (Precision Recall) représente la performance harmonique d'un système. 2

La précision, le rappel et le F-score de nos méthodes sont présentés dans le tableau X. Certaines catégories comme les protéines et l'ARN ont des performances bien supérieures à d'autres. La reconnaissance de la catégorie Protéine a le score F le plus élevé. Cette situation s'est également produite dans un autre système. Certains chercheurs pensent que cela est dû à un manque de données sur les trains. Les montants de chaque catégorie d'entité dans les données de formation sont présentés dans le tableau 1.

Cependant, après avoir comparé le tableau 1 et la figure 3, nous avons constaté que la quantité de données d'entraînement n'est pas la raison principale. La catégorie cell_type contient le petit ensemble de données d'entraînement, mais avec la plus haute précision et le deuxième score F le plus élevé. D'autre part, la catégorie ADN possède le deuxième plus grand ensemble de données de formation. La figure 4 montre que 12 % des mots de la catégorie « ADN-B » sont étiquetés à tort dans la catégorie « protéine B », ce qui est bien plus que les autres catégories biomédicales.

Tableau 1. Performance des principales catégories d'entités

Catégorie	rappel	précision	Score F
protéine	0,8062	0,6389	0,7129
cell_line	0,6160	0,5008	0,5525
type	0,6761	0,6427	0,6590
de cellule ADN	0,7356	0,7344	0,7350
ARN	0,6102	0,6050	0,6076
Dans l'ensemble	0,7610	0,6486	0,7003

[1] https://biopython.org/wiki/Main_Page
[2] https://code.google.com/p/word2vec/
[3] http://www.nactem.ac.uk/tsujii/GENIA/tagger/

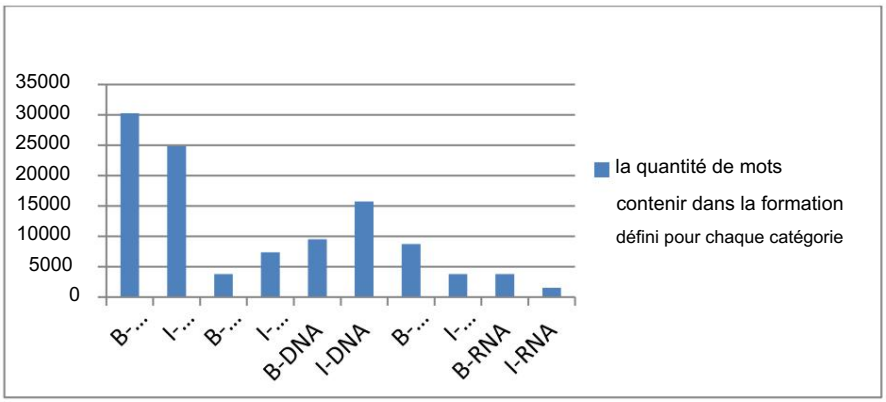

Figure 1. Données de formation pour les principales catégories d'entités

Après des recherches plus approfondies sur les données d'entraînement, nous avons trouvé deux raisons principales. La première est que les entités nommées biomédicales sont généralement composées de plusieurs entités nommées imbriquées. Pour les catégories d'ADN et de protéines, il existe de nombreuses entités nommées imbriquées qui se chevauchent. Pour Par exemple, il y a deux exemples de formation dans la figure 5. Les mots « Epstein-Barr », « virus », « EBV », « protéine », « cellule » et les signes de ponctuation « (», «) » sont intégrés dans les deux entités mais ils appartiennent à deux catégories différentes.

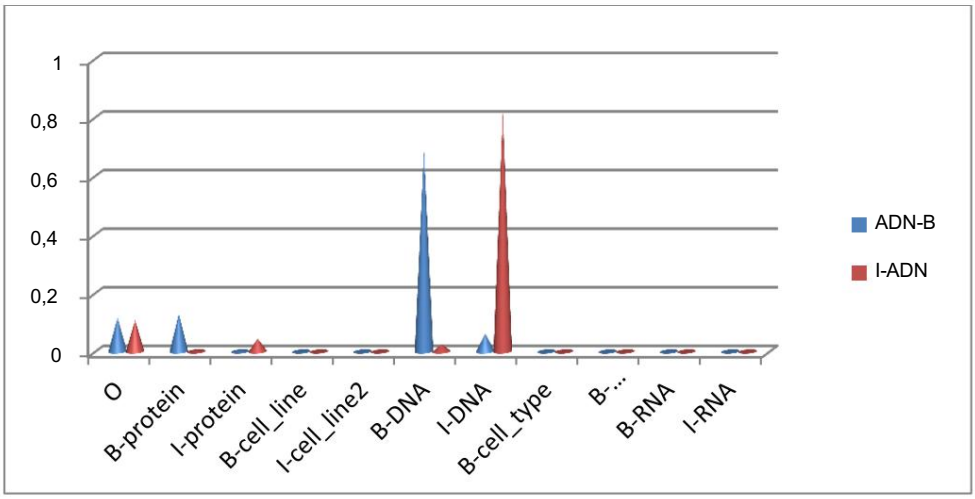

Figure 4. Répartition des erreurs

Après avoir analysé les données, nous constatons que ces mots peuvent apparaître à différentes positions selon différentes catégories. Par exemple, le mot « cellules » se situe généralement à la fin des entités de type cellulaire mais au milieu des entités de lignée cellulaire. Étant donné que la notation BIO actuelle ne peut pas présenter de telles informations, pour utiliser ce type d'informations, la notation BMESO est appliquée aux étapes intermédiaires du traitement.

NER1	Epstein-Barr	virus	(	EBV	)	latent	membrane	protéine	1	(	LMP-1	) oncogène
Étiquettes1	ADN-B	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN	I-ADN I-ADN	I-ADN
NER2	Epstein-Barr	virus	(	EBV	)	nucléaire	antigène	2	(	EBNA-2	)	protéine
Étiquettes2	Protéine B	I-protéine	I-protéine	I-protéine	I-protéine	I-protéine	I-protéine	I-protéine O		Protéine B O		Ô

Figure 5. Exemples de NER et d'étiquettes

La notation BMESO ressemble beaucoup à la notation BIO. Cependant, il peut donner une description plus détaillée de la position de chaque mot dans les entités. B indique toujours le début d'une entité. E indique la fin d'une entité. Les mots entre le mot de début et le mot de fin seront notés par M. Si l'entité est un seul mot, il sera marqué par S. Les expériences montrent que le BMESO peut améliorer les performances.

La deuxième raison est le manque d'ensemble de formation étiqueté et certaines entités du fichier de test n'apparaissent pas dans l'ensemble de formation. Dans le domaine biomédical, chaque jour, de nouveaux mots sont générés et certains disparaissent en même temps. Un tel cas est donc limité par la tâche elle-même et par les ensembles de données publiques.

Ensuite, nous obtenons les résultats finaux sur le fichier de test GENIA et les listons dans le tableau 2 suivant.

Tableau 2. Comparaisons avec certains systèmes de pointe

Équipes	rappel	précision	Score F
Notre résultat 76,13 Celui de		66,54	71.01
Sasaki et al.[28] 79,85 Celui de		68,58	73,78
Liao et al.[14] 73,6 Sun et al.		72,8	73.2
's[29] 72,3 ABNER[30] 72,0		70,2	71.2
Saha et al.'s[31] 67,66		69,1	70,5
		68,12	67,89

Par rapport à d'autres équipes, notre résultat est très proche des résultats de début d'art avec le moins de modèles de fonctionnalités et sans construire de lexique. L'approche basée sur le dictionnaire est bénéfique lorsque la composition du lexique est relativement stable. Cependant, le dictionnaire biomédical évolue chaque jour. Et le modèle de fonctionnalités de conception est un travail à forte intensité de main-d'œuvre, et il sera différent en raison de l'évolution des tâches et des corpus.
Les deux types de solutions ont une limite flexible.

4. Conclusion

Nous avons présenté un réseau neuronal à plusieurs couches et l'appliquons à la tâche biomédicale NER. Nous avons atteint des performances proches de l'état de l'art en utilisant le nouveau modèle. Les résultats révèlent que ces modèles de réseaux neuronaux sans fonctionnalités supplémentaires conçues à la main apportent d'excellentes performances.

Dans les travaux futurs, nous nous concentrerons sur d'autres moyens d'améliorer encore les performances du réseau neuronal. Premièrement, nous réalisons que le mot limite gauche est très important et peut affecter la reconnaissance d'une entité entière pouvant inclure plusieurs mots. Par conséquent, une fois que le premier mot est mal étiqueté, les mots suivants seront également incorrects. La liaison de reconnaissance inverse avec reconnaissance directe sera explorée ensuite. Deuxièmement, des fonctionnalités statistiques peuvent être intégrées.

Les références

[1] H. Dai, Y. Chang, R. Tzong-Han Tsai et W. Hsu, « Nouveaux défis pour l'exploration de textes biologiques au cours de la prochaine décennie », Journal of Computer Science and Technology, vol. 25, non. 1, (2010), pages 169-179.
[2] M. Krallinger, A. Morgan, L. Smith, F. Leitner, L. Tanabe, J. Wilbur, L. Hirschman et A. Valencia, « Évaluation des systèmes d'exploration de texte pour la biologie : aperçu de la deuxième biocréation Community Challenge", Biologie du génome, vol. 9, non. 2, (2008).
[3] H. Dai, C. Huang, R. Lin, R. Tsai et W. Hsu, "Recherche Web Biosmile : une application Web pour annoter les entités et les relations biomédicales", Nucleic Acids Research, vol. 36, (2008), pages 390 à 397.
[4] D. Rebholz-Schuhmann, M. Arregui, S. Gaudan, H. Kirsch et A. Jimeno, "Traitement de texte via les services Web : appeler Whatizit", Bioinformatics, vol. 24, non. 2, (2008), pages 296 à 300.
[5] L. Si, T. Kanungo et X. Huang, "Boosting Performance of Bio-Entity Recognition by Combining Results from Multiple Systems", Actes du 5ème atelier international sur la bioinformatique ACM, (2005), pp. 76-83 .
[6] A. Vlachos, « Évaluer et combiner les systèmes biomédicaux de reconnaissance d'entités nommées », BioNLP 2007 : Traitement du langage biologique, traductionnel et clinique, (2007), pp. 199-206.

[7] Y. Tsuruoka, Y. Tateishi, J.-D. Kim, T. Ohta, J. McNaught, S. Ananiadou et JI Tsujii, « Développement d'un marqueur de partie de discours robuste pour le texte biomédical, dans : Advances in Informatics », Springer, (2005), pp. 382-392.

[8] J. Finkel, S. Dingare, H. Nguyen, M. Nissim, C. Manning et G. Sinclair, « Exploiting Context for Biomedical Entity Recognition : From Syntax to the Web », Actes de l'atelier conjoint international sur les Traitement du langage en biomédecine et ses applications. Association pour la linguistique computationnelle, (2004), pp. 88-91.

[9] K. Lee, Y. Hwang, S. Kim et H. Rim, "Reconnaissance d'entités nommées biomédicales à l'aide d'un modèle en deux phases basé sur Svms", Journal of biomedical informatics, vol. 37, non. 6, (2004), pages 436-447.

[10] G. Zhou, J. Zhang, J. Su, D. Shen et C. Tan, « Reconnaître les noms dans les textes biomédicaux : une machine Approche d'apprentissage. Bioinformatique", vol. 20, n° 7, (2004), pp. 1178-1190.

[11] L. Li, R. Zhou et D. Huang, "Reconnaissance d'entités nommées biomédicales en deux phases à l'aide de Crfs. Biologie computationnelle et chimie", vol. 33, n° 4, (2009), pp. 334-338.

[12] C.-N. Hsu, Y.-M. Chang, C.-J. Kuo, Y.-S. Lin, H.-S. Huang et I.-F. Chung, "Intégration de modèles d'analyse bidirectionnelle haute dimension pour le marquage des mentions génétiques", Bioinformatics, vol. 24, non. 13, (2008), pages 286 à 294.

[13] RT Tsai, C.-L. Sung, H.-J. Dai, H.-C. Hung, T.-Y. Sung et W.-L. Hsu, "Nerbio : utilisation de conjonctions de mots sélectionnées, de normalisation de termes et de modèles globaux pour améliorer la reconnaissance biomédicale des entités nommées", BMC bioinformatics, vol. 7, non. 5, (2006).

[14] Z. Liao et H. Wu, "Biomedical Named Entity Recognition Based on Skip-Chain Crfs", Contrôle industriel et ingénierie électronique (ICICEE), Conférence internationale 2012 sur. IEEE, (2012), pages 1495 à 1498.

[15] L. Li, W. Fan et D. Huang, « Un système Bio-Ner en deux phases basé sur des classificateurs intégrés et une stratégie multi-agents » (2013).

[16] R. Collobert, « Deep Learning for Efficient Discriminative Parsing », Conférence internationale sur l'intelligence artificielle et les statistiques, (2011).

[17] RDY Bengio et P. Vincent, "Un modèle de langage probabiliste neuronal", dans NIPS, vol. 13, (2001).

[18] RCAJ Westion, « Une architecture unifiée pour le traitement du langage naturel : réseaux de neurones profonds avec apprentissage multitâche », dans ICML, vol. (2008).

[19] JWR Collobert, L. Bottou, M. Karlen, K, Kavukcuoglu et AP Kuksa, "Traitement du langage naturel (presque) à partir de zéro", JMLR, (2011).

[20] RED Yoshua Bengio, P. Vincent et C. Janvin, "Un modèle de langage probabiliste neuronal", J. Mach. Apprendre. Rés., vol. 3, (2003), pages 1137 à 1155.

[21] H. Schwenk, "Modèles de langage spatial continu", Computer Speech & Language, vol. 21, non. 3, (2007), pages 492 à 518.

[22] T. Mikolov, M. Karafiat, L. Burget, J. Cernocky et S. Khudanpur, "Recurrent Neural Network Based Language Model", onzième conférence annuelle de l'International Speech Communication Association (INTERSPEECH), (2010), pp . 1045-1048.

[23] A. Mnih et Y. W The, « Un algorithme rapide et simple pour la formation de modèles de langage probabilistes neuronaux », Actes de la 29e Conférence internationale sur l'apprentissage automatique (ICML-12), (2012), pp. 1751-1758.

[24] R. Collobert, « Deep Learning for Efficient Discriminative Parsing », In Conférence internationale sur Intelligence artificielle et statistiques (AISSTATS), (2011).

[25] LAR Joseph Turian et Y. Bengio, « Représentations de mots : une méthode simple et générale pour l'apprentissage semi-supervisé », Actes de la 48e réunion annuelle de l'Association for Computational Linguistics, (2010), pp. 384-394.

[26] WTY Tomas Mikolov et G. Zweig, "Linguistic Regularities in Continuous Space Word Representations", Actes de la conférence 2013 de la section nord-américaine de l'Association for Computational Linguistics: Human Language Technologies, (2013), pp. 746- 751.

[27] L. Bottou, « Apprentissage par gradient stochastique dans les réseaux de neurones », Actes de Neuro-Nîmes, vol. 91, (1991).

[28] YTY Sasaki, J. McNaught et S. Ananiadou, "Comment tirer le meilleur parti des dictionnaires Ne dans Statistical Ner. Proc. Atelier Tendances actuelles dans le traitement biomédical du langage naturel", (2008), pp. 63-70.

[29] YGCH Sun, XL Wang et L. Lin, "Champs aléatoires conditionnels basés sur des fonctionnalités riches pour la reconnaissance d'entités nommées biologiques", Computers in Biology and Medicine, vol. 37, (2007), p. 1327-1333.

[30] Abner : Un système de reconnaissance d'entités nommées biomédicales, (2013), pp. 46-51.

[31] SNSK Saha, S. Sarkar et P. Mitra, "Un noyau composite pour la reconnaissance d'entités nommées", Pattern Recognition Letters, vol. 3, (2010), pages 1591-1597.

